**Michael Rose**
11/21/11
Data Mining

# Project 8: Mushroom Data Exploration

## Data

The first step in using this data was to preprocess the CSV datafile and add column headings to each for ease of use in KNIME. Additionally, Weka will explode without column heading. In Weka, the summary statistics were relatively useful in determining the usefulness of various attributes. Veil-type and veil-color are nearly worthless. Veil-types are all 'partial', and veil-color was almost all white with a few yellow, orange, and brown.

## Classification

The first attempt at classification, a Naïve Bayes classifier, worked relatively well achieving 95.8% accuracy. After generating association rules, classification was relatively simple to improve upon the accuracy of the classifier. Noting the extremely high correlations of association rules, the Ripper (JRip) algorithm was used to build a classifier with 100% accuracy in classifying the dataset.

## Clustering

Clustering this dataset given all the ordinal data is relatively difficult. Ideally, we could create some sort of taxonomy through hierarchal clustering. Without sampling the data, it was relatively infeasible to run a hierarchal clustering. The Weka implementation of the hierarchal clusterer ate past 4GB of JVM Heapsize. Simple K-means with Euclidean Distance or Manhattan Distance both returned 41%/59% for each cluster, unfortunately with 37.62% of instances misclassified. DBScan ran for several minutes with an eps=2, minPoints=6 and created 16 clusters. Clustering is somewhat of a dead-end for this dataset.

## Association Analysis

When generating rules, it was noted that the confidences of the rules found grew with the number of rules found, showing that there were many good rules of high confidence not found in the initial rule generation of k=10 rules, k=100 rules. With a rule generation of 10,000, all were found above 0.97 confidence. While many rules may not be worthwhile, this is a good indication of being able to successfully use a rule-based classifier.

## Anomaly Detection

Density-based anomaly detection doesn't seem to find any noise, or when it does, there are nearly 21 clusters. This doesn't necessarily imply that there are no anomalies. A statistical approach is nearly impossible, Gaussian distributions have little meaning in this dataset due to the lack of continuous values.

**Can you generate summary statistics that help describe the data?**

The only meaningful summary statistics that can be generated are the counts and percentages of each attribute. Very few attributes, like perhaps ring-number or gill-size could perhaps be degenerated into continuous values, but given the homogenous data would mean relatively little.

**Can the edible and poisonous data objects be distilled into groups?**

Yes. The poisonous mushrooms can be distilled into groups extraordinarily easily using association analysis and using a number of rules in order to classify the data. JRip has 100% success in classifying these into groups.

**Can a classification model be created that can predict whether a mushroom is edible or poisonous?**

Yes, a rule-based classifier has extremely high success in classifying a mushroom as edible or poisonous. This is due to the high confidence/support rules that are able to be generated, JRip generates a classifier capable of classifying the entire dataset with 100% accuracy.

**Do any anomalies exist in the dataset?**

Unknown. Though, according to the metadata those mushrooms of unknown edibility have been folded into the poisonous class. When classifying, these could account for some of the anomalously classified instances.

**Can any association rules be generated from this dataset?**

Any number of association rules can be generated, when generating 10-100 rules. The dataset lends its self well to rule generation, 10000 rules can be generated of high confidence (0.97+).